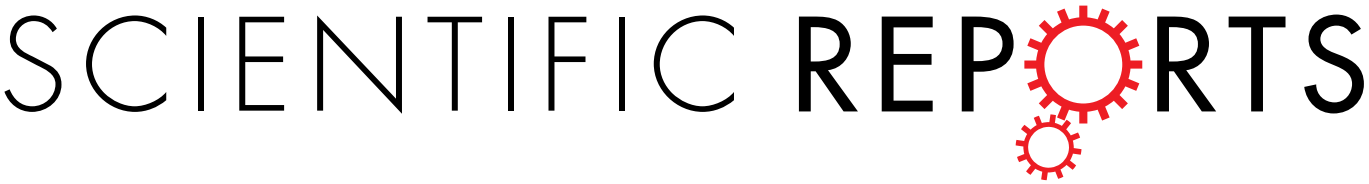

OPEN

A taxonomic signature of obesity in a large study of American adults

Brandilyn A. Peters[1], Jean A. Shapiro[2], Timothy R. Church[3], George Miller[4,5,6], Chau Trinh-Shevrin[1,6], Elizabeth Yuen[7], Charles Friedlander[7], Richard B. Hayes[1,6] & Jiyoung Ahn[1,6]

Received: 16 February 2018
Accepted: 12 June 2018
Published online: 27 June 2018

Animal models suggest that gut microbiota contribute to obesity; however, a consistent taxonomic signature of obesity has yet to be identified in humans. We examined whether a taxonomic signature of obesity is present across two independent study populations. We assessed gut microbiome from stool for 599 adults, by 16S rRNA gene sequencing. We compared gut microbiome diversity, overall composition, and individual taxon abundance for obese (BMI $\geq$ 30 kg/m^2), overweight (25 $\leq$ BMI $<$ 30), and healthy-weight participants (18.5 $\leq$ BMI $<$ 25). We found that gut species richness was reduced ($p = 0.04$), and overall composition altered ($p = 0.04$), in obese (but not overweight) compared to healthy-weight participants. Obesity was characterized by increased abundance of class Bacilli and its families Streptococcaceae and Lactobacillaceae, and decreased abundance of several groups within class Clostridia, including Christensenellaceae, Clostridiaceae, and Dehalobacteriaceae ($q < 0.05$). These findings were consistent across two independent study populations. When random forest models were trained on one population and tested on the other as well as a previously published dataset, accuracy of obesity prediction was good (~70%). Our large study identified a strong and consistent taxonomic signature of obesity. Though our study is cross-sectional and causality cannot be determined, identification of microbes associated with obesity can potentially provide targets for obesity prevention and treatment.

The World Health Organization estimates that global obesity prevalence has more than doubled since 1980, classifying >600 million adults as obese in 2014. Obesity increases risk for many diseases, including cancer, atherosclerosis, and diabetes[1–3]. While the fundamental cause of obesity is an imbalance between energy intake and expenditure, other factors may modify susceptibility, such as genetics[4], epigenetics[5], and gut microbial composition[6]. Because of the potential to modify bacterial communities, the microbiome is an enticing candidate to target for obesity prevention and treatment. Reaching this goal requires identification of specific taxa and/or microbial functions associated with obesity in humans; once identified, further downstream experimentation can establish whether these taxa and/or functions are causative agents[7], and, if so, suggest interventions.

Experiments in germ-free mice colonized with gut microbiota from wild-type mice[8], obese mice[9], or obese humans[10], demonstrate that microbiota play a critical role in adiposity in test systems. Moreover, these experiments have demonstrated transmissibility of obese phenotypes via gut microbes. These findings lead to the question of whether gut microbial composition confers susceptibility to obesity in humans. An early report in a small human sample ($n = 14$)[11] was consistent with findings in mice that obesity, whether genetic[12] or diet-induced[13,14], is associated with an increase in relative abundance of the Firmicutes phylum, and a decrease in relative abundance of the Bacteroidetes phylum. However, more recent studies in humans have not corroborated this pattern[15–21]. Recent meta-analyses of studies with 16S rRNA gene data have not found consistent obesity-related taxonomic signatures across studies[22–24]. Small sample sizes, heterogeneous populations, insufficient confounder control, and different methodologies may contribute to disagreement between studies.

Using data from two independent cross-sectional studies of older American adults ($n = 599$), we aimed to: (1) examine whether within-person microbial diversity (α-diversity) and between-person differences in overall microbial composition (β-diversity) are associated with obesity, and (2) identify specific taxa and inferred

[1]Department of Population Health, New York University School of Medicine, New York, NY, USA. [2]Division of Cancer Prevention and Control, Centers for Disease Control and Prevention, Atlanta, GA, USA. [3]Division of Environmental Health Sciences, School of Public Health, University of Minnesota, Minneapolis, MN, USA. [4]Department of Surgery, New York University School of Medicine, New York, NY, USA. [5]Department of Cell Biology, New York University School of Medicine, New York, NY, USA. [6]NYU Perlmutter Cancer Center, New York University School of Medicine, New York, NY, USA. [7]Kips Bay Endoscopy Center, New York, NY, USA. Correspondence and requests for materials should be addressed to J.A. (email: Jiyoung.Ahn@nyumc.org)

metagenomic functions associated with obesity. The latter aim may provide targets for research on obesity treatment and prevention.

Results

Participant characteristics. Descriptive characteristics of healthy-weight, overweight, and obese participants are presented in Table 1. Participants were initially recruited for a colonoscopy-screening study, and approximately half (48%) had asymptomatic colorectal polyps detected at study screening or a previous screening. Participants were predominantly white (94%) and above middle-age (62 ± 7 years old). The overweight and obese groups had higher percentages of men than the healthy-weight group ($p < 0.0001$), while race and age distributions did not differ significantly across BMI categories. Data on energy intake and exercise were available in the New York University (NYU) study only. Daily energy intake did not differ significantly across BMI categories, although a weak positive correlation was detected between energy intake and continuous BMI (Spearman $r = 0.17$, $p = 0.02$). Additionally, overweight and obese participants exercised less frequently than healthy-weight participants ($p = 0.01$).

α- and β-diversity in relation to obesity. Globally, BMI category was associated with richness (i.e. number of OTUs) ($p = 0.002$) and the Shannon index ($p = 0.03$), but not with evenness ($p = 0.14$), at a rarefaction depth of 1,490 sequence reads/sample (Supplemental Table 1). In pairwise comparisons, richness was reduced in obese compared to healthy-weight participants ($b = -9.87$, $p = 0.04$, $p_{Holm} = 0.08$); this pattern was apparent, though not statistically significant, for the Shannon index ($b = -0.11$, $p = 0.11$, $p_{Holm} = 0.22$) and evenness ($b = -0.01$, $p = 0.22$, $p_{Holm} = 0.44$) (Fig. 1a–c; Supplemental Table 1). Overweight participants did not differ significantly from healthy-weight participants for any of these α-diversity indices (Supplemental Table 1). Partial constrained analysis of principal coordinates (CAP) of the weighted UniFrac distance revealed separation of obese from both healthy-weight and overweight participants on the main axis, with overweight separated from healthy-weight participants on the secondary axis (Fig. 1e), although principal coordinate analysis (PCoA) did not reveal clustering by BMI category (Fig. 1d). In permutational multivariate analysis of variance (PERMANOVA) analysis of the weighted UniFrac distance, BMI category was not associated globally with overall microbiome composition ($p = 0.14$). In pairwise comparisons, overall microbiome composition differed between obese and healthy-weight participants ($p = 0.04$, $p_{Holm} = 0.07$), while overweight and healthy-weight participants did not differ significantly ($p = 0.64$, $p_{Holm} = 0.64$) (Supplemental Table 1). When further classifying obese participants as class I ($30 < BMI \leq 35\,kg/m^2$; $n = 90$) or class II-III ($BMI > 35\,kg/m^2$; $n = 52$), we observed that both classes of obesity tended to differ from healthy-weight participants in richness and overall microbiome composition, though not with statistical significance (Supplemental Fig. 1; Supplemental Table 1).

The relationship of obesity with overall microbiome diversity and composition was consistent in both the Centers for Disease Control and Prevention (CDC) and NYU studies, and in those with and without asymptomatic colorectal polyps (Supplemental Fig. 2a,b; Supplemental Table 2). We observed a significant reduction in richness in obese vs. healthy-weight women ($p = 0.03$), however this was not observed in men ($p = 0.47$) (Supplemental Fig. 2a; Supplemental Table 2). In the NYU study, availability of diet ($n = 171$) and exercise ($n = 175$) data allowed us to assess whether exercise or intake of total energy, fiber, fat, or protein confounded the association of obesity with microbiome diversity and composition. We observed that adjustment for these variables did not attenuate differences in diversity and composition between obese and healthy-weight participants in the NYU study (Supplemental Table 3).

Taxa associated with obesity. We examined differential abundance of taxa by BMI at the phylum through OTU levels (Supplemental Table 4). Contrary to several previous reports, abundances of the two most prevalent phyla, Firmicutes and Bacteroidetes, were not associated with BMI category ($p = 0.40$ and $p = 0.49$, respectively). The Firmicutes/Bacteroidetes ratio was also not associated with BMI category (Kruskal-Wallis test $p = 0.94$). However, several sub-taxa within Firmicutes were associated with obesity. The Bacilli class (fold change [FC] = 2.93) and its Streptococcaceae (FC = 2.42), Lactobacillaceae (FC = 6.23), and Gemellaceae (FC = 2.3) families were elevated in obese compared to healthy-weight participants. Within class Clostridia, the Christensenellaceae (FC = 0.57), Clostridiaceae (FC = 0.58), Dehalobacteriaceae (FC = 0.34), and SHA-98 (FC = 0.49) families were depleted, and the Veillonellaceae family enriched (FC = 1.46), in obese compared to healthy-weight participants. Greater abundances of family Actinomycetaceae of phylum Actinobacteria, and family Enterobacteriaceae of phylum Proteobacteria, were also noted in obese participants, as were decreased abundances of family Rikenellaceae (Bacteroidetes phylum) and Pasteurellaceae (Proteobacteria phylum) (Fig. 2; Supplemental Table 4). Similar to findings in obese participants, overweight participants had increased abundance of Lactobacillaceae and Streptococcaceae, and decreased abundance of Christensenellaceae, Clostridiaceae, and Dehalobacteriaceae, compared to healthy-weight participants (Fig. 2; Supplemental Table 4).

At OTU level, 90 OTUs were identified as differentially abundant globally by BMI category at $q < 0.05$ (Fig. 3; Supplemental Table 4). OTUs in *Streptococcus* and Proteobacteria (Enterobacteriaceae and *Bilophila*) were enriched in obese compared to healthy-weight participants. Within Clostridia, several patterns emerged when comparing obese to healthy-weight participants, including enrichment of *Blautia* OTUs, and depletion of *Coprococcus*, *Oscillospira*, Clostridiaceae, Christensenellaceae, and *Dehalobacterium* OTUs, in the obese. Additionally, many unclassified OTUs within Clostridia (Ruminococcaceae and unclassified families) were depleted in the obese. Fewer OTUs were differentially abundant between overweight and healthy-weight participants, though findings were similar to those in obese participants (Supplemental Table 4).

When stratifying these analyses by sex, we observed some similarities between men and women (Supplemental Table 5). For example, obese men and women both had increased Bacilli, *Streptococcus*, and Gammaproteobacteria, and decreased Christensenellaceae, Clostridiaceae, and Dehalobacteriaceae, than healthy-weight men and women, respectively (though not always reaching $p_{Holm} < 0.05$).

	Healthy-weight	Overweight	Obese	p^b
Combined (n = 599)	n = 211	n = 246	n = 142	
Men (%)	37.9	69.5	49.3	<0.0001
Age (y; mean ± SD)	62.7 ± 7.7	62.1 ± 7.0	61.7 ± 6.1	0.32
Race (%)				0.26
White	95.3	93.9	93.7	
Black	1.4	3.3	4.2	
Other	3.3	2.0	0.7	
Missing	0	0.8	1.4	
Colorectal polyps[c] (%)	42.7	50.8	51.4	0.15
BMI (kg/m^2; mean ± SD)	22.6 ± 1.7	27.1 ± 1.4	35.0 ± 5.0	<0.0001
CDC (n = 423)	n = 130	n = 173	n = 120	
Men (%)	35.4	68.8	49.2	<0.0001
Age (y; mean ± SD)	62.8 ± 4.7	62.2 ± 5.1	62.4 ± 4.8	0.50
Race (%)				0.68
White	96.9	97.7	96.7	
Black	1.5	0.6	2.5	
Other	1.5	1.7	0.8	
Missing	0	0	0	
Colorectal polyps (%)	34.6	44.5	50.0	0.04
BMI (kg/m^2; mean ± SD)	22.7 ± 1.6	27.1 ± 1.4	34.9 ± 5.0	<0.0001
NYU (n = 176)	n = 81	n = 73	n = 22	
Men (%)	42.0	71.2	50.0	0.001
Age (y; mean ± SD)	62.4 ± 10.8	61.8 ± 10.2	57.7 ± 9.9	0.18
Race (%)				0.06
White	92.6	84.9	77.3	
Black	1.2	9.6	13.6	
Other	6.2	2.7	0	
Missing	0	2.8	9.1	
Colorectal polyps (%)	55.6	65.8	59.1	0.43
BMI (kg/m^2; mean ± SD)	22.3 ± 1.8	27.0 ± 1.4	35.5 ± 5.4	<0.0001
Daily energy intake[d,e] (kcal; mean ± SD)	1,703 ± 755	1,846 ± 723	1,830 ± 719	0.34
Exercise[d] (%)				0.01
None	7.4	16.4	27.3	
<1 hr/week	7.4	8.2	27.3	
1 hr/week	11.1	8.2	0	
2 hr/week	11.1	13.7	18.2	
3 hr/week	29.6	24.7	4.5	
4+ hr/week	33.3	28.8	18.2	
Missing	0	0	4.5	

Table 1. Characteristics of participants in the CDC and NYU studies by BMI[a]. [a]Healthy-weight: $18.5 \leq$ BMI < 25 kg/m^2; Overweight: $25 \leq$ BMI < 30 kg/m^2; Obese: BMI ≥ 30 kg/m^2. [b]P-value for difference between BMI categories from Kruskal-Wallis test for continuous variables and X^2 test for categorical variables. [c]Had one or more colorectal polyps currently or previously identified. [d]Variable only available in NYU study (n = 171 for energy intake, n = 175 for exercise). [e]Determined from food frequency questionnaire.

Inferred metagenome pathways associated with obesity. The KEGG pathway "alpha-Linolenic acid (ALA) metabolism" was differentially abundant globally by BMI category ($q < 0.05$); in pairwise comparisons, this pathway was enriched in obese compared to healthy-weight participants ($p < 0.0001$, $p_{Holm} < 0.0001$) (Supplemental Table 6). We also investigated whether several *a priori* pathways, related to hypothesized mechanisms of microbial involvement in obesity (discussed later), were nominally associated with obesity (Supplemental Table 6). "Butanoate (butyrate) metabolism" was marginally depleted ($p = 0.06$, $p_{Holm} = 0.11$), while "secondary bile acid biosynthesis" was marginally enriched ($p = 0.08$, $p_{Holm} = 0.17$), in obese compared to healthy-weight participants. "Lipopolysaccharide biosynthesis", "propanoate (propionate) metabolism", and "methane metabolism" were not associated with obesity. Interestingly, several families depleted in obese compared to healthy-weight participants (Christensenellaceae, Clostridiaceae, Dehalobacteriaceae, and SHA-98) were positively associated with butanoate and propanoate metabolism, and inversely associated with secondary bile acid biosynthesis (Fig. 4). We also explored whether OTUs associated with obesity contributed to abundance of KEGG orthologs for butyrate synthesis genes, butyrate kinase and butyryl-CoA:acetate CoA transferase[25].

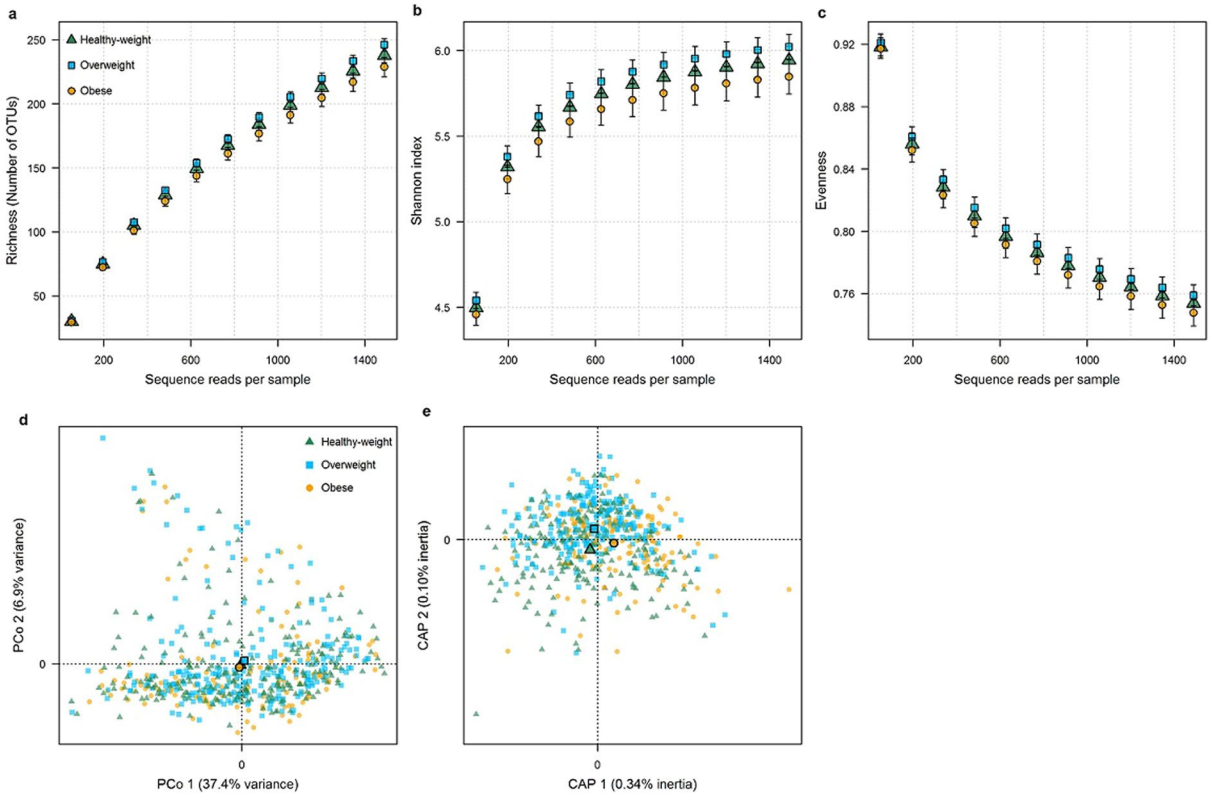

Figure 1. α-diversity and β-diversity in relation to BMI. (**a**–**c**) Richness, Shannon diversity index, and Evenness rarefaction curves in healthy-weight, overweight, and obese participants. Rarefaction curves were estimated by taking the mean of the α-diversity indices averaged for each participant over 100 iterations at each rarefaction sequencing depth. (**d**) Principal coordinate analysis of the weighted UniFrac distances. Shapes outlined in black represent centroids for healthy-weight, overweight, and obese participants. (**e**) Partial constrained analysis of principal coordinates (CAP) based on the weighted UniFrac distance. BMI category was the constraining variable, and sex, age, polyp status, and study were conditioning variables.

While several obesity-depleted OTUs did contribute to butyrate synthesis KEGG orthologs (e.g. OTUs from Christensenellaceae, *Oscillospira*, *SMB53*, Clostridiales, Rikenellaceae), obesity-enriched OTUs also contributed to these orthologs (Supplemental Fig. 3).

Homogeneity of results across two independent populations. We observed consistencies in taxa associated globally with BMI category ($q < 0.05$) between the CDC and NYU studies (Supplemental Table 7; Fig. 2), despite the much smaller sample size of the NYU study. In pairwise comparisons in both studies, obese participants had increased abundance of Bacilli (Streptococcaceae and Lactobacillaceae families) and Gammaproteobacteria, and decreased abundance of Christensenellaceae, compared to healthy-weight participants ($p_{Holm} < 0.05$). At the OTU level, we observed substantially more OTUs associated globally with BMI category ($q < 0.05$) in the CDC study than in the NYU study, likely due to the substantially smaller sample size of the NYU study, and the large number of tests. We therefore explored similarities between the studies at the OTU level using nominal p-values. 17 OTUs were associated with obesity ($p < 0.05$) in the same direction in both studies, while only 2 OTUs were associated with obesity ($p < 0.05$) in the opposite direction between the studies (Supplemental Table 8; Fig. 5). The OTUs overlapping across the studies in significance and direction included Gemellaceae, *Streptococcus*, and *Blautia* OTUs (increased in the obese), and *Parabacteroides*, Clostridiaceae, Lachnospiraceae, Ruminococcaceae, Clostridiales, and *Oscillospira* OTUs (decreased in the obese).

Microbiome-based classification of obesity. We generated a random forest model based on 1,825 OTUs in the CDC study (training set) to predict obesity in the NYU and Baxter *et al.*[26] studies (testing sets). We used the area under the curve-random forest (AUC-RF) algorithm to perform a backward elimination process based on the initial ranking of OTUs in a random forest model; this algorithm identifies the optimal random forest model (and optimal set of predictive OTUs) as the model with the highest AUC. Our optimal model included 49 OTUs and had an AUC of 0.81 (Fig. 6). We then performed repeated cross-validation of the AUC-RF process to more accurately determine the model's predictive accuracy; the mean AUC from repeated cross-validation was 0.65. We used the Youden's index of the ROC curve as the probability threshold above which a subject was classified as

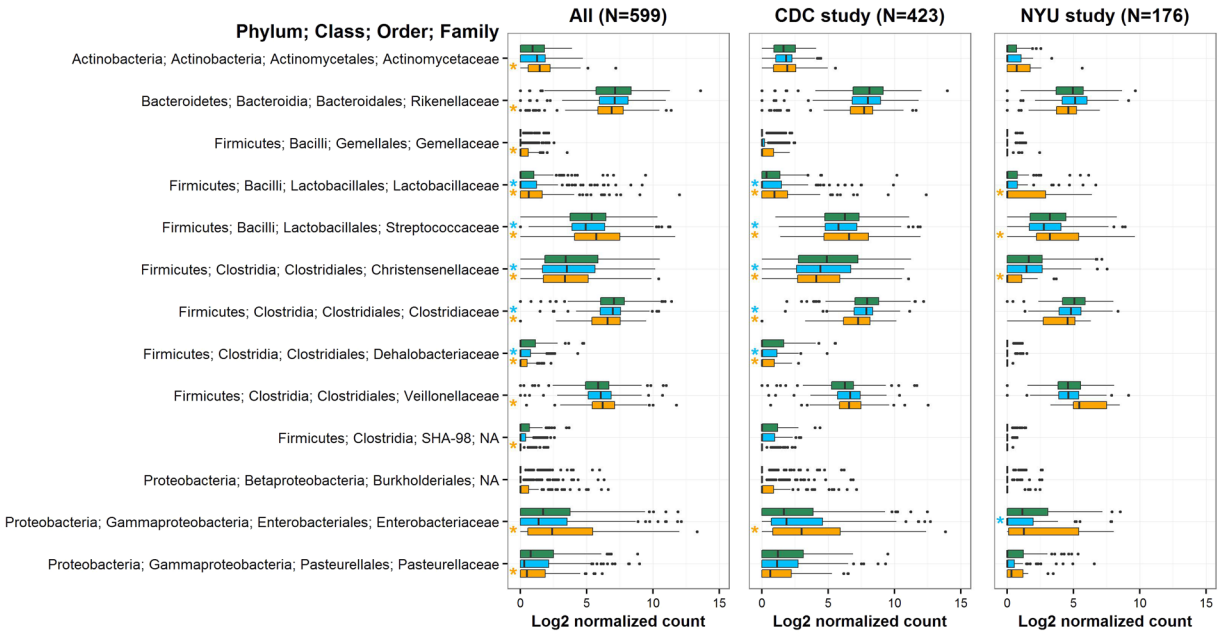

Figure 2. Count boxplots of families that were differentially abundant by obesity. Families associated globally with BMI category in the DESeq2 analysis (LRT $q < 0.05$) were included in the plot. Green, blue, and orange boxplots represent healthy-weight, overweight, and obese participants, respectively. Counts were normalized for DESeq2 size factors and log2 transformed after adding a pseudocount of 1. Stars to the left-hand side of boxplots indicate significant difference in abundance from healthy-weight ($p_{Holm} < 0.05$ for pairwise comparison).

obese in the testing sets. The accuracy of the model in correctly classifying subjects as obese or non-obese when applied to the NYU and Baxter *et al*. testing sets was 0.72 and 0.68, respectively.

Discussion

In this large study of older American adults, we observed that obesity was associated with reduced gut microbial richness and alterations in overall gut microbial composition. These findings point to a possible effect of gut microbial composition on energy balance or storage. The homogeneity of our results in two independent study populations, and the good accuracy of obesity classification with a microbiome-based machine learning model, reveals an emerging taxonomic signature of obesity which may have implications for obesity prevention and treatment.

Several mechanisms have been hypothesized through which gut bacteria may affect host energy balance or storage. The "energy harvest" hypothesis posits that bacteria contribute to obesity by extracting energy from otherwise indigestible dietary fiber, through production of digestible short-chain fatty acids (SCFAs)[9]. The "metabolic endotoxemia" hypothesis posits that plasma lipopolysaccharide (LPS, or endotoxin) derived from the cell wall of Gram-negative bacteria elicits low-grade inflammation, promoting adiposity[27,28]. A final broad category of mechanisms is that of microbial metabolites or products modulating energy balance[7]. Notably, SCFAs, in addition to being energy sources to the host, are important signaling molecules with beneficial effects for host energy metabolism[29], and protect against diet-induced obesity in animal models[30,31]. Other bacterial metabolites, such as methane[32] and secondary bile acids[33], may also modulate host energy balance. Here, we observed many taxonomic composition alterations associated with obesity. Whether and by what mechanism these bacterial groups impact obesity remains unclear, but we discuss some potential mechanisms in relation to our findings below.

Decreases in putative SCFA-producing bacteria in the obese may lend support to the hypothesis that SCFAs beneficially modulate host energy metabolism. The Christensenellaceae family is known to produce SCFAs, primarily acetate and butyrate[34], and was identified as the most heritable taxon in a study of 416 twin pairs; in that study, Christensenellaceae, Dehalobacteriaceae, SHA-98, Methanobacteriaceae, RF39, and *Oscillospira* were depleted in obese subjects compared to healthy-weight[35], much in agreement with our findings. Higher Christensenellaceae abundance in mice that received human fecal transplants was correlated with reduced weight gain, and transplant of obese donor stool amended with *Christensenella minuta* to recipient mice led to reduced adiposity[35]. Findings of depleted Christensenellaceae in obese individuals have since been replicated in other large studies[21,36], indicating that Christensenellaceae may be important for promoting leanness. *Oscillospira* has also been suggested to promote human leanness; it was enriched in healthy-weight subjects in several human studies[36,37], and may contribute to leanness by degrading host glycans and producing SCFAs[37]. We also observed that other Clostridiales OTUs (Ruminococcaceae, Lachnospiraceae, and unclassified families) were depleted in the obese; although functions of these bacteria are unknown, many members of these families produce SCFAs[25]. An important caveat is that multitudes of gut bacteria produce SCFAs, making it unclear whether this mechanism is actually responsible for patterns observed. Our inferred metagenome analysis, however, revealed that the

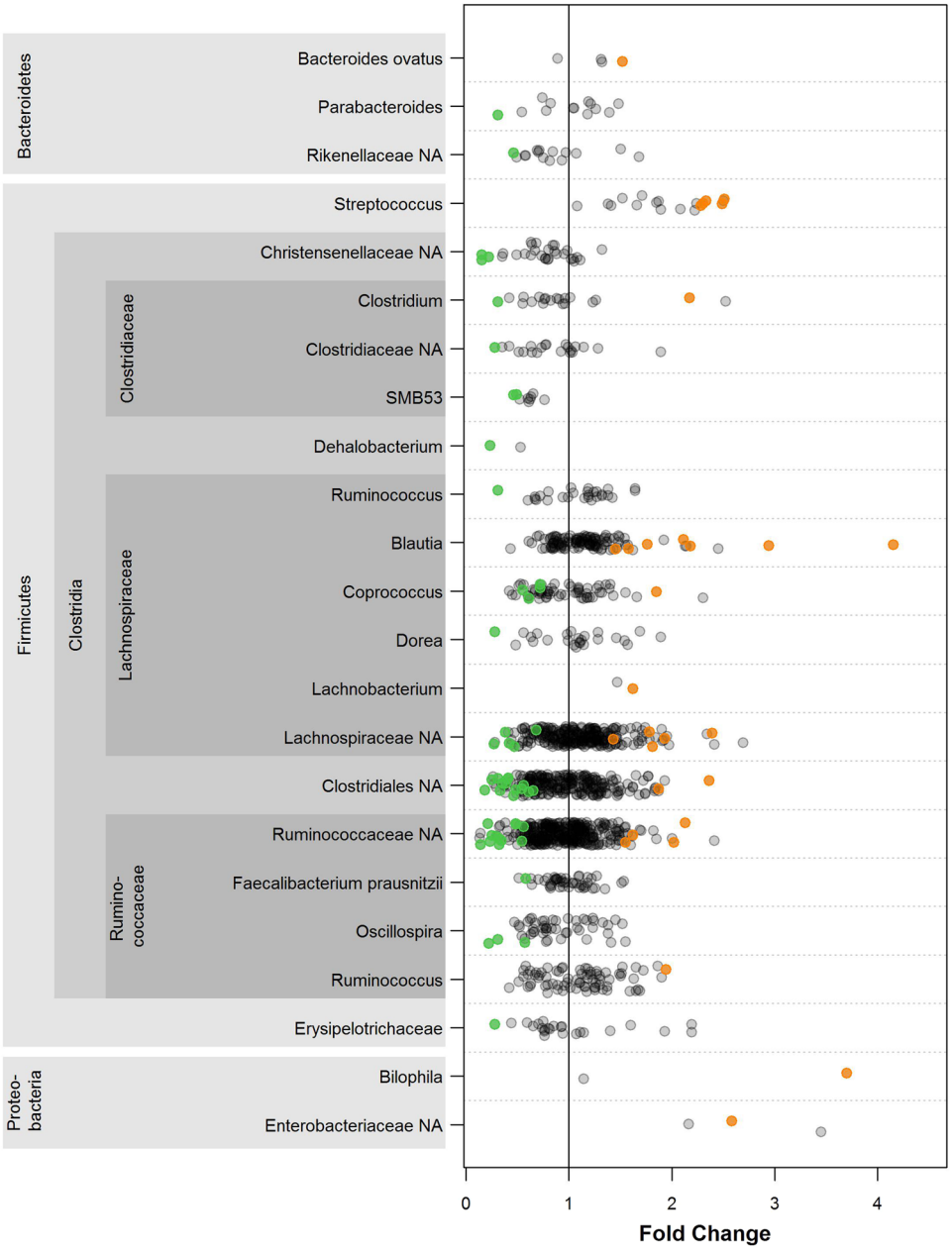

Figure 3. OTUs associated with obesity. OTU fold changes for obese vs. healthy-weight comparison in DESeq2 analysis are plotted. All OTUs within the given taxonomic groups are plotted, and orange and green points represent OTUs significantly ($p_{Holm} < 0.05$) higher or lower in abundance, respectively, in obese compared to healthy-weight participants. Only taxonomic groups with at least one differentially abundant OTU ($p_{Holm} < 0.05$) are displayed. "NA" indicates a group that was unclassified at the family, genus, or species level.

KEGG pathway related to the SCFA butyrate ("butanoate metabolism") was marginally depleted in obese compared to healthy-weight participants, supporting the beneficial SCFA hypothesis.

Increases in Enterobacteriaceae in the obese may lend support to the "metabolic endotoxemia" hypothesis, as LPS from Enterobacteriaceae exhibits high endotoxin activity[27]; however the "LPS biosynthesis" pathway was not associated with obesity in our inferred metagenome analysis. Enterobacteriaceae species were also associated with obesity in other studies[38,39], and have been shown to decrease following weight-loss interventions[40,41].

Synthesis of secondary bile acids and methane represent other potential mechanisms by which gut microbiota may modulate host energy balance. In our inferred metagenome analysis, we observed that "secondary bile acid biosynthesis" was marginally enriched in obese compared to healthy-weight participants, while "methane metabolism" was not associated with obesity. Some species in *Clostridium* and *Eubacterium* generate secondary bile acids[42], which may modulate adiposity via farnesoid X receptor (FXR) or Takeda G-protein-coupled receptor 5 (TGR5) signaling[33,43]. Methanogens may promote adiposity via conversion of hydrogen to methane gas[32,44],

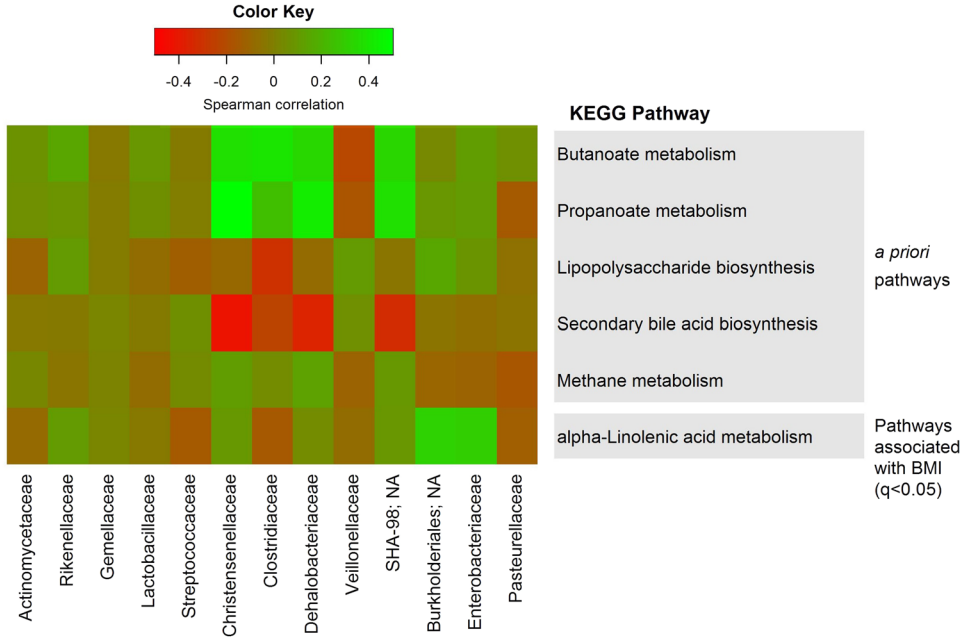

Figure 4. Correlations of bacterial families and inferred metagenomic functions. Family and KEGG pathway counts were DESeq2-normalized. Partial Spearman's correlation coefficients were estimated for each pairwise comparison of family and KEGG pathway abundance, adjusting for age, sex, study, and polyp status. KEGG pathways included in the heatmap were identified *a priori* or were associated globally with BMI category (LRT $q < 0.05$); families included in the heatmap were associated globally with BMI category (LRT $q < 0.05$).

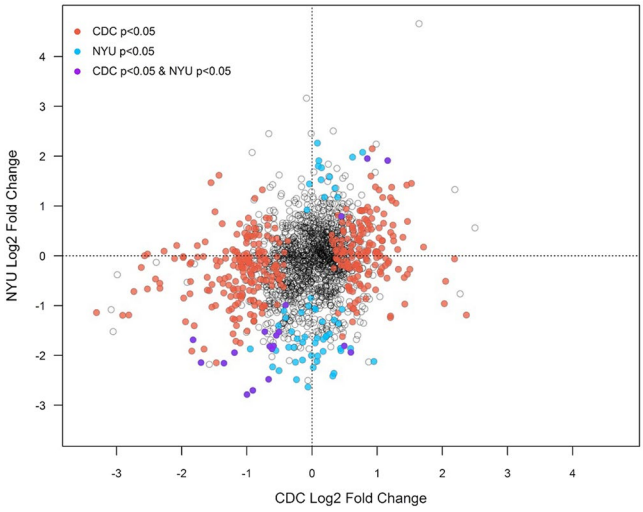

Figure 5. Scatterplot of obesity-associated OTUs in the CDC and NYU studies. All of the OTUs tested (1,825) are plotted by their log2 fold changes (obese vs. healthy-weight) in the CDC and NYU studies. OTUs represented by black open circles were not significantly associated with obesity in either study. Red, blue, and purple circles represent OTUs associated with obesity ($p < 0.05$) in the CDC study only, NYU study only, or in both studies, respectively. OTU models with extreme outliers (maximum Cook's distance >15) are not colored in the plot. $R^2 = 4.8\%$.

and have previously been associated with leanness by other studies[16,35,45], or, in contrast, with obesity[46–48]. More research is needed in human populations to elucidate the roles of secondary bile acids and methane in obesity.

We also identified that the "ALA metabolism" KEGG pathway was enriched in obese compared to healthy-weight participants. ALA is a type of n-3 polyunsaturated fatty acid, which may be metabolized to conjugated linolenic acids by gut microbiota[49]; conjugated linolenic acids were shown to have anti-adipogenic properties in several studies[50], in contradiction with this observed result.

We observed reduced microbial diversity in the obese, particularly for women. Obesity-related reductions in microbial diversity have been reported previously[15,21,36,39,51], though not by all[22,52]. One study related the reduction

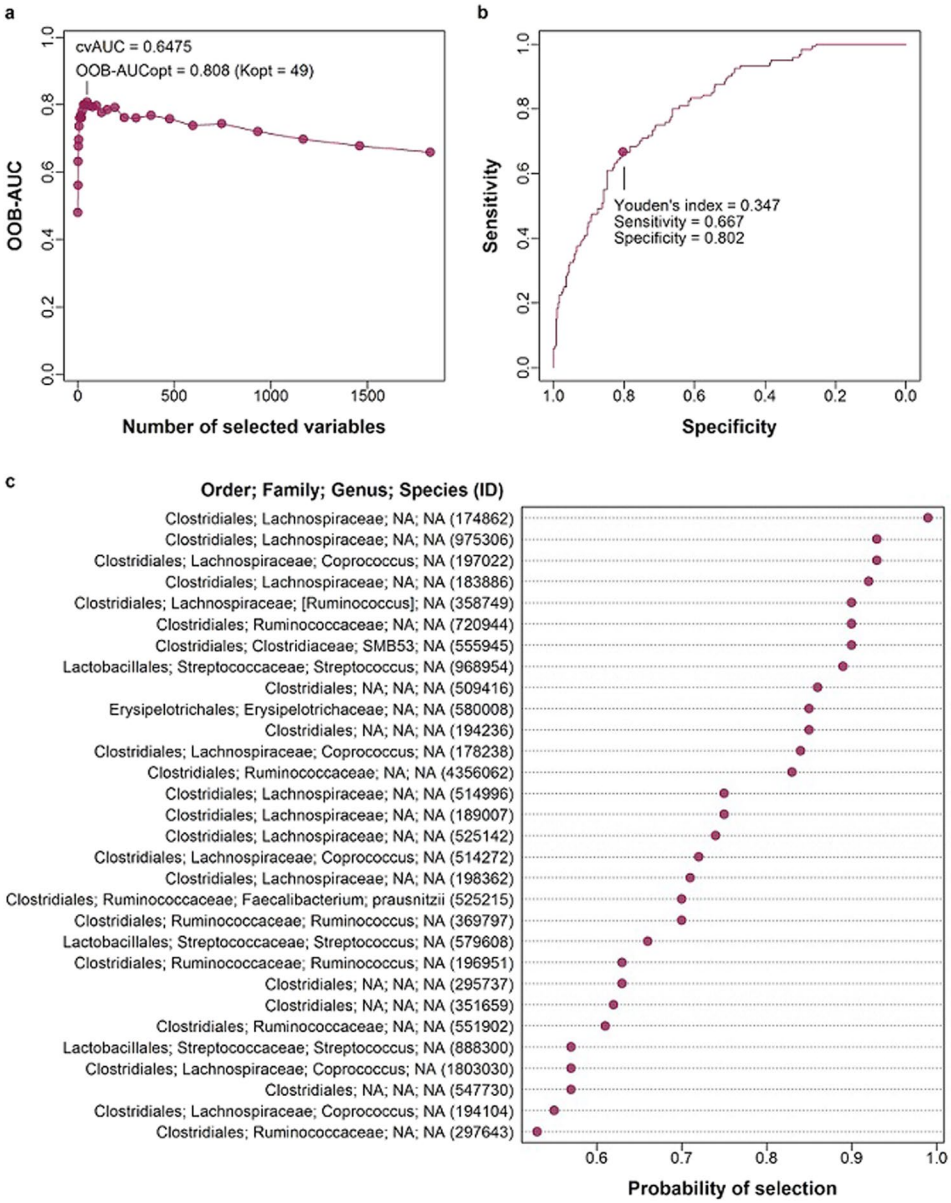

Figure 6. Random forest model of the training data set (CDC study). A random forest model was generated based on 1,825 DESeq2-normalized OTUs in the training data set (CDC study) using the AUCRF R package. (**a**) The optimal random forest model of 49 OTUs was selected by optimizing the area under the receiver operating characteristic (ROC) curve (AUC) of the random forest (optimal AUC = 0.81); the mean AUC of repeated (20 times) 5-fold cross-validations of the random forest model was 0.65. (**b**) ROC curve of the optimal random forest model, highlighting Youden's index (probability at maximum sum of sensitivity and specificity). (**c**) Top 30 OTUs with highest probability of selection in repeated cross-validation of the optimal random forest model.

in diversity to "abnormal energy input" in obesity[15]. Individuals with low microbial gene richness are more likely to be obese and have poorer metabolic health[53]. Additionally, a weight-loss intervention was less effective at improving inflammatory markers in those with low microbial gene richness[54]. Therefore, low microbial diversity may be a further factor conferring susceptibility to obesity. The reason for the sex difference in our microbiome diversity result is unclear; a possible mechanism may lie in the effect of sex hormones on the gut microbiota[55], however replication of the result in other studies is warranted.

The potential for manipulation of gut microbiota has generated interest in identifying a taxonomic signature for obesity that is responsible for the obesogenic mechanisms detailed above. Animal studies and some small human studies have demonstrated that the obese microbiome is characterized by a phylum-level signature of increased Firmicutes and decreased Bacteroidetes[8,9,12,15]. However, larger human studies have failed to replicate this signature[21–23], including the current study. It is possible that in humans, the taxonomic signature of obesity exists on a finer species (OTU) level, rather than at phylum level. Further, due to large between-person and between-population variability in the gut microbiome, large sample sizes are likely needed to detect such a

signature. This signature may differ by population factors such as age, race, and geography. We have observed consistency of findings between our two independent study populations, which both consisted of older, mostly white Americans, suggesting that a taxonomic signature of obesity can be identified within homogeneous populations. In support of this, a recent meta-analysis robustly replicated eight obesity-associated OTUs across three large population-based cohorts of European descent[36]. Additionally, we observed good accuracy (~70%) of obesity classification by a microbiome-based random forest model, trained on one study and tested on two studies with similar population characteristics to the training set. An analysis of 10 published datasets by another group observed overall poor accuracy of random forest models trained on one dataset and tested on the other nine (median accuracy 33–65% for 10 models)[24]. However these datasets differed substantially on population characteristics such as age, race, and geography, which all may impact model performance; here we have focused on homogeneous populations, assuming there is no universal taxonomic signature of obesity across all populations. Additionally, the authors used genus-level information to develop the models, whereas here we used OTU-level information, which could also impact model performance. Regardless of whether high accuracy of obesity classification can be achieved with machine learning, it remains possible that specific taxa play a mechanistic role in obesity.

Strengths of this study include the large sample size, control of potential confounders, comprehensive bacterial profiling, and availability of dietary data in a subset of participants. The effect of diet on gut microbial composition has been demonstrated previously[56–59]; due to effects of diet on both microbiota and BMI, it is difficult to tease apart potential microbial contributions to obesity from effects of diet on microbiota. Here, adjustment for dietary factors did not impact the association of obesity with microbial composition. Although power of this analysis was limited due to the small subset with dietary information (n = 171) and measurement error inherent in food frequency questionnaires[60], it suggests a relationship between microbial composition and obesity independent of diet. Our study also has several limitations. The cross-sectional design does not allow us to establish temporality or causality of the microbiome-obesity relationship. Additionally, due to the older age and mostly white study population (96% 50 and over; 94% white), findings may not be generalizable to younger or more diverse populations. We also lacked antibiotic usage information in the CDC study which did not allow us to exclude individuals taking antibiotics, and we lacked dietary and exercise data in the CDC study which did not allow us to adjust for these potential confounders in the full study population. Finally, lack of shotgun-sequenced metagenome data did not allow us to actually characterize metagenomic functions.

In summary, in this large study of older American adults, we observed a significant relationship between the gut microbiome and obesity. The taxa identified may open new avenues for experimental research on causal microbial agents of obesity. Additional large-scale studies are warranted in humans to confirm a taxonomic signature of obesity (in a variety of populations, as the signature may vary by age, race, and geography). From there, interventions in animals and humans can identify obesity-promoting bacteria or lean-promoting bacteria, and the mechanisms of their action. Looking forward, precision medicine approaches based on an individual's microbiome may eventually be used to effectively treat or prevent obesity.

Methods

Study population. We included data from two independent study populations based at colonoscopy clinics: the Centers for Disease Control and Prevention Study of In-home Tests for Colorectal Cancer (CDC study)[61], and the New York University Human Microbiome and Colorectal Tumor study (NYU study)[62] (Supplemental Fig. 4). The CDC study was approved by the institutional review boards of University of Minnesota and the CDC, and the NYU study by the institutional review board of NYU School of Medicine. Methods were carried out in accordance with relevant guidelines and regulations, and all participants provided informed consent.

The CDC study contributed 451 subjects at University of Minnesota/Minnesota Gastroenterology (12/2012-7/2014). Eligible participants were 50–75 years old, scheduled to have a colonoscopy for routine screening, able to read English, and not currently taking anticoagulants. Additionally, participants must not have had >1 episode of rectal bleeding in the last six months, a positive FOBT in the past twelve months, a colonoscopy in the past 5 years, a personal history of colorectal cancer, polyps, or inflammatory bowel disease, or a personal or family history of familial adenomatous polyposis or hereditary nonpolyposis colorectal cancer. We excluded participants that withdrew (n = 17), subjects for whom sequencing failed (n = 4), subjects missing BMI (n = 3), and underweight subjects (BMI < 18.5 kg/m^2; n = 4), resulting in 423 subjects.

The NYU study enrolled 239 participants from Kips Bay Endoscopy Center in New York City (6/2012-8/2014). Eligible participants were 18 years or older (range: 29–86), recently underwent colonoscopy, able to read English, and not on long-term antibiotics. We excluded participants missing colonoscopy reports (n = 2), missing BMI (n = 9), or underweight (n = 1), and further excluded participants with rectal bleeding (n = 18) or with personal history of colorectal cancer (n = 10), inflammatory bowel disease (n = 22), anastomosis (n = 6), or familial adenomatous polyposis (n = 1), in order to conform the NYU study to the CDC study; exclusion based on these non-mutually exclusive criteria resulted in 176 subjects.

Stool samples. Subjects collected stool onto Beckman Coulter Hemoccult II SENSA® cards (Beckman Coulter, CA) at home. This method produces reproducible and accurate 16S rRNA gene-derived microbiota data[63,64], and exhibits stability at room temperature up to 8 weeks[65]. CDC samples were mailed to a laboratory for fecal occult blood testing within several days of stool collection; this testing does not impact microbiota composition[62,63]. After testing, CDC samples were refrigerated at 4 °C until shipment to NYU, and upon arrival were stored at −80 °C (range: 7–183 days from sample collection to receipt by NYU). NYU samples were mailed directly to NYU following at-home collection and stored immediately at −80 °C.

Microbiome assay. DNA was extracted from stool using the PowerLyzer PowerSoil Kit (Mo Bio Laboratory Inc., CA) following manufacturer's protocol, as described previously[62]. Barcoded amplicons were generated covering the V4 region of the 16S rRNA gene using the F515/R806 primer pair[66]. The PCR reaction, using FastStart High Fidelity PCR system, dNTP pack (Roche, IN), was run as follows: initial denaturing at 94 °C for 3 min, followed by 25 cycles of 94 °C for 15 s, 52 °C for 45 s and 72 °C for 1 min, and a final extension at 72 °C for 8 min. PCR products were purified using Agencourt AMPure XP (Beckman Coulter Life Sciences, IN) and quantified using the Agilent 4200 TapeStation (Agilent Technologies, CA). Amplicon libraries were pooled at equal molar concentrations and sequenced on Illumina MiSeq with a 300-cycle (2×151 bp) kit.

Sequence read processing. Forward and reverse reads were joined using *join_paired_ends.py* in QIIME with default parameters[67]. Sequences were demultiplexed, and poor-quality sequences excluded, using default parameters of QIIME script *split_libraries_fastq.py*; median sequence length was 253 base pairs. Chimeric sequences were excluded using USEARCH 6.1, with the "gold" reference database (Broad Institute Microbiome Utilities microbiomeutil-r20110519). Sequence reads were clustered into operational taxonomic units (OTUs) against the Greengenes 13_8 reference sequence collection, using QIIME *pick_closed_reference_otus.py* script (results were highly similar using *de novo* OTU picking, data not shown). The final dataset of 599 participants included 15,098,120 sequence reads (mean ± SD: 25,206 ± 15,616 reads/sample) and 8,902 OTUs. Quality control data showing excellent reproducibility for this data has been published previously[62].

Covariates. Only limited demographic information (age, sex, BMI, race) was collected during CDC study enrollment. The NYU study collected more extensive information (e.g. data on exercise, smoking, health history, and dental health) and food frequency questionnaires. The food frequency questionnaire used in the NYU study was the 137-item DQX from the National Cancer Institute Prostate, Lung, Colorectal, and Ovarian Cancer screening trial (PLCO), available at https://biometry.nci.nih.gov/cdas/datasets/plco/97/. Nutrient variables were calculated following the PLCO protocol; briefly, the frequency for each line item was multiplied by a nutrient amount (derived from the USDA CSFII database) which was dependent on the gender of the subject as well as the response to serving size, when applicable. Healthy-weight was defined as BMI $\geq$ 18.5 and $<$25 kg/m^2, overweight as BMI $\geq$ 25 and $<$30 kg/m^2, and obese as BMI $\geq$ 30 kg/m^2. Colorectal polyps were identified at colonoscopy and confirmed by pathology; cases were defined as those with $\geq$1 polyp of non-normal histology, or those with history of polyps.

α-diversity. α-diversity (within-subject species diversity) was assessed using richness, Shannon diversity index, and evenness, calculated in 100 iterations for rarefied OTU tables (minimum: 50 reads/sample, maximum: 1,490 reads/sample [lowest participant sequencing depth]) using QIIME script *alpha_rarefaction.py*. We examined whether α-diversity (at 1,490 sequence reads/sample) differed across BMI categories using linear regression, adjusting for age, sex, polyp status, and study. Statistical significance of the global BMI category variable was determined using an F-test comparing the full vs. reduced model (i.e. without BMI category). P-values for the two pairwise comparisons of interest (obese vs. healthy-weight and overweight vs. healthy-weight) were adjusted with the Holm method[68].

β-diversity. β-diversity (between-subject species diversity) was assessed using the weighted UniFrac distance[69]. Principal coordinate analysis (PCoA)[70] and partial constrained analysis of principal coordinates (CAP)[71] were used to visually explore the relationship between BMI and overall bacterial composition. In partial CAP analysis, BMI category was the constraining variable, and sex, age, polyp status, and study were conditioning variables. Permutational multivariate analysis of variance (PERMANOVA)[72] was used to examine statistically whether overall bacterial composition differed by BMI category, adjusting for age, sex, polyp status, and study. Statistical significance was determined as described above for α-diversity.

Differential abundance testing. To examine differences in abundance of bacterial taxa across BMI categories we used negative binomial generalized linear models (DESeq2)[73]. This method models raw counts with a negative binomial distribution, adjusting internally for "size factors" which normalize for differences in sequencing depth between samples. The raw counts of 8,902 OTUs were agglomerated to 14 phyla, 30 classes, 56 orders, 115 families, 302 genera, and 413 species. Prior to analysis, we filtered the data to include only taxa with $\geq$2 sequence reads in $\geq$5% of participants (30 participants), resulting in inclusion of 11 phyla, 20 classes, 25 orders, 52 families, 100 genera, 133 species, and 1,825 OTUs. DESeq2 models were adjusted for age, sex, polyp status, and study. DESeq2 default outlier replacement, independent filtering of low-count taxa, and filtering of count outliers were turned off. We used likelihood-ratio tests (LRT) to determine statistical significance of the global BMI category variable in DESeq2 models; we adjusted the p-values for taxa at each level (i.e. class, genus) for the false discovery rate (FDR)[74], with models with maximum Cook's distance $>$15 removed prior to p-value adjustment. For models that were significant (LRT FDR-adjusted p-value [q-value] $<$ 0.05), Wald test p-values for the two pairwise comparisons of interest (obese vs. healthy-weight and overweight vs. healthy weight) were adjusted with the Holm method[68]. This methodology controls the mixed directional FDR[75].

Inferred metagenomes. PiCRUST[76] was used to infer metagenomic content from 16S rRNA gene-based microbial compositions. The 5,753 observed KEGG[77] gene orthologs were grouped into 276 KEGG pathways. We filtered the data to include only pathways with $\geq$2 reads in $\geq$30 participants, and removed unclassified pathways and pathways related to "Human Diseases" or "Organismal Systems", resulting in inclusion of 185 pathways. We used DESeq2 (as described above) to test differences in pathway abundance across BMI categories. Statistical significance was determined as described above for differential abundance testing. We considered nominal p-values

for *a priori* pathways of interest, and q-values for other pathways. We used partial Spearman's correlations to examine associations between taxa and pathways, adjusting for age, sex, study, and polyp status. We also explored OTU contributions to *a priori* KEGG orthologs of interest using PiCRUST script *metagenome_contributions.py*.

Random forest machine learning. We used a random forest model based on the CDC study (training set) to classify individuals in the NYU study and another human study[26] (testing sets) as obese (BMI $\geq 30\,kg/m^2$) or non-obese (BMI $< 30\,kg/m^2$). We chose the Baxter *et al*. study due to its similarity with our study, as it was also colonoscopy-based and comprised of older, mostly white Americans. The Baxter *et al*. data was downloaded from the NCBI Sequence Read Archive (SRP062005) and processed identically to our data (see "Sequence read processing" in Methods), to facilitate comparison with our studies. After excluding participants with cancer, the Baxter *et al*. data comprised 402 subjects (age mean $\pm$ SD $= 59.5 \pm 11.7$, 91% white, 50% men). The random forest model for the training set was generated using the AUCRF R package[78], which performs variable selection based on optimizing the area under the receiver operating characteristic (ROC) curve (AUC) of the random forest. DESeq2-normalized counts of 1,825 OTUs were used in variable selection. We performed repeated (20 times) 5-fold cross-validation of the random forest model. The probability threshold above which a subject was classified as obese in the testing sets was based on Youden's index (probability at maximum sum of sensitivity and specificity) of the ROC curve of the training set model. Accuracy was calculated as (true positives + true negatives)/(total subjects).

Diet and exercise sensitivity analysis. In the NYU study, data on diet (e.g. total energy, fiber, protein, fat intake) and exercise were available, and we checked whether adjusting for these variables in the NYU study influenced our overall (α- and β-) diversity results. Models with fiber, protein, or fat intake were adjusted for total energy. Those with unrealistic total energy intake (<500 or >4000 kcal/day; $n = 3$) and those leaving blank $>50\%$ of the items on the 137-item food frequency questionnaire ($n = 2$) were considered missing and excluded from the dietary analysis. Those missing exercise data ($n = 1$) were excluded from the exercise analysis.

Data availability statement. The datasets analyzed during the current study are available in the dbGaP repository (accession phs001381.v1.p1).

References

1. Guh, D. P. *et al*. The incidence of co-morbidities related to obesity and overweight: a systematic review and meta-analysis. *BMC public health* **9**, 88, https://doi.org/10.1186/1471-2458-9-88 (2009).
2. Renehan, A. G., Tyson, M., Egger, M., Heller, R. F. & Zwahlen, M. Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *Lancet (London, England)* **371**, 569–578, https://doi.org/10.1016/s0140-6736(08)60269-x (2008).
3. Bogers, R. P. *et al*. Association of overweight with increased risk of coronary heart disease partly independent of blood pressure and cholesterol levels: a meta-analysis of 21 cohort studies including more than 300 000 persons. *Archives of internal medicine* **167**, 1720–1728, https://doi.org/10.1001/archinte.167.16.1720 (2007).
4. Lu, Y. & Loos, R. J. Obesity genomics: assessing the transferability of susceptibility loci across diverse populations. *Genome medicine* **5**, 55, https://doi.org/10.1186/gm459 (2013).
5. Waterland, R. A. Epigenetic mechanisms affecting regulation of energy balance: many questions, few answers. *Annual review of nutrition* **34**, 337–355, https://doi.org/10.1146/annurev-nutr-071813-105315 (2014).
6. Ley, R. E. Obesity and the human microbiome. *Current opinion in gastroenterology* **26**, 5–11, https://doi.org/10.1097/MOG.0b013e328333d751 (2010).
7. Harley, I. T. & Karp, C. L. Obesity and the gut microbiome: Striving for causality. *Molecular metabolism* **1**, 21–31, https://doi.org/10.1016/j.molmet.2012.07.002 (2012).
8. Backhed, F. *et al*. The gut microbiota as an environmental factor that regulates fat storage. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 15718–15723, https://doi.org/10.1073/pnas.0407076101 (2004).
9. Turnbaugh, P. J. *et al*. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027–1031, https://doi.org/10.1038/nature05414 (2006).
10. Ridaura, V. K. *et al*. Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science (New York, N.Y.)* **341**, 1241214, https://doi.org/10.1126/science.1241214 (2013).
11. Ley, R. E., Turnbaugh, P. J., Klein, S. & Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **444**, 1022–1023, https://doi.org/10.1038/4441022a (2006).
12. Ley, R. E. *et al*. Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 11070–11075, https://doi.org/10.1073/pnas.0504978102 (2005).
13. Turnbaugh, P. J., Backhed, F., Fulton, L. & Gordon, J. I. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell host & microbe* **3**, 213–223, https://doi.org/10.1016/j.chom.2008.02.015 (2008).
14. Murphy, E. F. *et al*. Composition and energy harvesting capacity of the gut microbiota: relationship to diet, obesity and time in mouse models. *Gut* **59**, 1635–1642, https://doi.org/10.1136/gut.2010.215665 (2010).
15. Turnbaugh, P. J. *et al*. A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484, https://doi.org/10.1038/nature07540 (2009).
16. Armougom, F., Henry, M., Vialettes, B., Raccah, D. & Raoult, D. Monitoring bacterial community of human gut microbiota reveals an increase in Lactobacillus in obese patients and Methanogens in anorexic patients. *PloS one* **4**, e7125, https://doi.org/10.1371/journal.pone.0007125 (2009).
17. Schwiertz, A. *et al*. Microbiota and SCFA in lean and overweight healthy subjects. *Obesity (Silver Spring, Md.)* **18**, 190–195, https://doi.org/10.1038/oby.2009.167 (2010).
18. Mai, V., McCrary, Q. M., Sinha, R. & Glei, M. Associations between dietary habits and body mass index with gut microbiota composition and fecal water genotoxicity: an observational study in African American and Caucasian American volunteers. *Nutrition journal* **8**, 49, https://doi.org/10.1186/1475-2891-8-49 (2009).
19. Duncan, S. H. *et al*. Human colonic microbiota associated with diet, obesity and weight loss. *International journal of obesity (2005)* **32**, 1720–1724, https://doi.org/10.1038/ijo.2008.155 (2008).
20. Tims, S. *et al*. Microbiota conservation and BMI signatures in adult monozygotic twins. *Isme j* **7**, 707–717, https://doi.org/10.1038/ismej.2012.146 (2013).

21. Yun, Y. *et al*. Comparative analysis of gut microbiota associated with body mass index in a large Korean cohort. *BMC microbiology* **17**, 151, https://doi.org/10.1186/s12866-017-1052-0 (2017).
22. Walters, W. A., Xu, Z. & Knight, R. Meta-analyses of human gut microbes associated with obesity and IBD. *FEBS letters* **588**, 4223–4233, https://doi.org/10.1016/j.febslet.2014.09.039 (2014).
23. Finucane, M. M., Sharpton, T. J., Laurent, T. J. & Pollard, K. S. A taxonomic signature of obesity in the microbiome? Getting to the guts of the matter. *PloS one* **9**, e84689, https://doi.org/10.1371/journal.pone.0084689 (2014).
24. Sze, M. A. & Schloss, P. D. Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. *mBio* **7**, https://doi.org/10.1128/mBio.01018-16 (2016).
25. Vital, M., Howe, A. C. & Tiedje, J. M. Revealing the bacterial butyrate synthesis pathways by analyzing (meta)genomic data. *mBio* **5**, e00889, https://doi.org/10.1128/mBio.00889-14 (2014).
26. Baxter, N. T., Ruffin, M. T. T., Rogers, M. A. & Schloss, P. D. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome medicine* **8**, 37, https://doi.org/10.1186/s13073-016-0290-3 (2016).
27. Zhao, L. The gut microbiota and obesity: from correlation to causality. *Nature reviews. Microbiology* **11**, 639–647, https://doi.org/10.1038/nrmicro3089 (2013).
28. Cani, P. D. *et al*. Changes in gut microbiota control metabolic endotoxemia-induced inflammation in high-fat diet-induced obesity and diabetes in mice. *Diabetes* **57**, 1470–1481, https://doi.org/10.2337/db07-1403 (2008).
29. Kimura, I., Inoue, D., Hirano, K. & Tsujimoto, G. The SCFA Receptor GPR43 and EnergyMetabolism. *Frontiers in endocrinology* **5**, 85, https://doi.org/10.3389/fendo.2014.00085 (2014).
30. Lin, H. V. *et al*. Butyrate and propionate protect against diet-induced obesity and regulate gut hormones via free fatty acid receptor 3-independent mechanisms. *PloS one* **7**, e35240, https://doi.org/10.1371/journal.pone.0035240 (2012).
31. den Besten, G. *et al*. Short-Chain Fatty Acids Protect Against High-Fat Diet-Induced Obesity via a PPARgamma-Dependent Switch From Lipogenesis to Fat Oxidation. *Diabetes* **64**, 2398–2408, https://doi.org/10.2337/db14-1213 (2015).
32. Mathur, R. *et al*. Intestinal Methanobrevibacter smithii but not total bacteria is related to diet-induced weight gain in rats. *Obesity (Silver Spring, Md.)* **21**, 748–754, https://doi.org/10.1002/oby.20277 (2013).
33. Parséus, A. *et al*. Microbiota-induced obesity requires farnesoid X receptor. *Gut* **66**, 429–437, https://doi.org/10.1136/gutjnl-2015-310283 (2017).
34. Morotomi, M., Nagai, F. & Watanabe, Y. Description of Christensenella minuta gen. nov., sp. nov., isolated from human faeces, which forms a distinct branch in the order Clostridiales, and proposal of Christensenellaceae fam. nov. *International journal of systematic and evolutionary microbiology* **62**, 144–149, https://doi.org/10.1099/ijs.0.026989-0 (2012).
35. Goodrich, J. K. *et al*. Human genetics shape the gut microbiome. *Cell* **159**, 789–799, https://doi.org/10.1016/j.cell.2014.09.053 (2014).
36. Beaumont, M. *et al*. Heritable components of the human fecal microbiome are associated with visceral fat. *Genome biology* **17**, 189, https://doi.org/10.1186/s13059-016-1052-7 (2016).
37. Konikoff, T. & Gophna, U. Oscillospira: a Central, Enigmatic Component of the Human Gut Microbiota. *Trends in microbiology*, https://doi.org/10.1016/j.tim.2016.02.015 (2016).
38. Yasir, M. *et al*. Comparison of the gut microbiota of people in France and Saudi Arabia. *Nutrition & diabetes* **5**, e153, https://doi.org/10.1038/nutd.2015.3 (2015).
39. Verdam, F. J. *et al*. Human intestinal microbiota composition is associated with local and systemic inflammation in obesity. *Obesity (Silver Spring, Md.)* **21**, E607–615, https://doi.org/10.1002/oby.20466 (2013).
40. Xiao, S. *et al*. A gut microbiota-targeted dietary intervention for amelioration of chronic inflammation underlying metabolic syndrome. *FEMS microbiology ecology* **87**, 357–367, https://doi.org/10.1111/1574-6941.12228 (2014).
41. Sotos, M. *et al*. Gut microbes and obesity in adolescents. *The Proceedings of the Nutrition Society* **67**, 1–E20, https://doi.org/10.1017/S0029665108006290 (2008).
42. Wahlstrom, A., Sayin, S. I., Marschall, H. U. & Backhed, F. Intestinal Crosstalk between Bile Acids and Microbiota and Its Impact on Host Metabolism. *Cell metabolism* **24**, 41–50, https://doi.org/10.1016/j.cmet.2016.05.005 (2016).
43. Watanabe, M. *et al*. Bile acids induce energy expenditure by promoting intracellular thyroid hormone activation. *Nature* **439**, 484–489, http://www.nature.com/nature/journal/v439/n7075/suppinfo/nature04330_S1.html (2006).
44. Pimentel, M., Gunsalus, R. P., Rao, S. S. C. & Zhang, H. Methanogens in Human Health and Disease. *Am J Gastroenterol Suppl* **1**, 28–33 (2012).
45. Million, M. *et al*. Correlation between body mass index and gut concentrations of Lactobacillus reuteri, Bifidobacterium animalis, Methanobrevibacter smithii and Escherichia coli. *International journal of obesity (2005)* **37**, 1460–1466, https://doi.org/10.1038/ijo.2013.20 (2013).
46. Mathur, R. *et al*. Methane and hydrogen positivity on breath test is associated with greater body mass index and body fat. *The Journal of clinical endocrinology and metabolism* **98**, E698–702, https://doi.org/10.1210/jc.2012-3144 (2013).
47. Mbakwa, C. A. *et al*. Gut colonization with methanobrevibacter smithii is associated with childhood weight development. *Obesity (Silver Spring, Md.)* **23**, 2508–2516, https://doi.org/10.1002/oby.21266 (2015).
48. Lee, H. S. *et al*. Associations among organochlorine pesticides, Methanobacteriales, and obesity in Korean women. *PloS one* **6**, e27773, https://doi.org/10.1371/journal.pone.0027773 (2011).
49. Druart, C. *et al*. Role of the lower and upper intestine in the production and absorption of gut microbiota-derived PUFA metabolites. *PloS one* **9**, e87560, https://doi.org/10.1371/journal.pone.0087560 (2014).
50. Hennessy, A. A., Ross, P. R., Fitzgerald, G. F. & Stanton, C. Sources and Bioactive Properties of Conjugated Dietary Fatty Acids. *Lipids* **51**, 377–397, https://doi.org/10.1007/s11745-016-4135-z (2016).
51. Dominianni, C. *et al*. Sex, body mass index, and dietary fiber intake influence the human gut microbiome. *PloS one* **10**, e0124599, https://doi.org/10.1371/journal.pone.0124599 (2015).
52. Kasai, C. *et al*. Comparison of the gut microbiota composition between obese and non-obese individuals in a Japanese population, as analyzed by terminal restriction fragment length polymorphism and next-generation sequencing. *BMC gastroenterology* **15**, 100, https://doi.org/10.1186/s12876-015-0330-2 (2015).
53. Le Chatelier, E. *et al*. Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546, https://doi.org/10.1038/nature12506 (2013).
54. Cotillard, A. *et al*. Dietary intervention impact on gut microbial gene richness. *Nature* **500**, 585–588, https://doi.org/10.1038/nature12480 (2013).
55. Moreno-Indias, I. *et al*. Neonatal Androgen Exposure Causes Persistent Gut Microbiota Dysbiosis Related to Metabolic Disease in Adult Female Rats. *Endocrinology* **157**, 4888–4898, https://doi.org/10.1210/en.2016-1317 (2016).
56. Xu, Z. & Knight, R. Dietary effects on human gut microbiome diversity. *The British journal of nutrition* **113**(Suppl), S1–5, https://doi.org/10.1017/s0007114514004127 (2015).
57. David, L. A. *et al*. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563, https://doi.org/10.1038/nature12820 (2014).
58. Turnbaugh, P. J. *et al*. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. *Science translational medicine* **1**, 6ra14, https://doi.org/10.1126/scitranslmed.3000322 (2009).
59. Wu, G. D. *et al*. Linking long-term dietary patterns with gut microbial enterotypes. *Science (New York, N.Y.)* **334**, 105–108, https://doi.org/10.1126/science.1208344 (2011).

60. Thompson, F. E. *et al*. The National Cancer Institute's Dietary Assessment Primer: A Resource for Diet Research. *Journal of the Academy of Nutrition and Dietetics* **115**, 1986–1995, https://doi.org/10.1016/j.jand.2015.08.016 (2015).
61. Shapiro, J. A. *et al*. A Comparison of Fecal Immunochemical and High-Sensitivity Guaiac Tests for Colorectal Cancer Screening. *The American journal of gastroenterology* **112**, 1728–1735, https://doi.org/10.1038/ajg.2017.285 (2017).
62. Peters, B. A. *et al*. The gut microbiota in conventional and serrated precursors of colorectal cancer. *Microbiome* **4**, 69, https://doi.org/10.1186/s40168-016-0218-6 (2016).
63. Sinha, R. *et al*. Collecting Fecal Samples for Microbiome Analyses in Epidemiology Studies. *Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **25**, 407–416, https://doi.org/10.1158/1055-9965.epi-15-0951 (2016).
64. Dominianni, C., Wu, J., Hayes, R. B. & Ahn, J. Comparison of methods for fecal microbiome biospecimen collection. *BMC microbiology* **14**, 103, https://doi.org/10.1186/1471-2180-14-103 (2014).
65. Song, S. J. *et al*. Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies. *mSystems* **1**, https://doi.org/10.1128/mSystems.00021-16 (2016).
66. Caporaso, J. G. *et al*. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America* **108**(Suppl 1), 4516–4522, https://doi.org/10.1073/pnas.1000080107 (2011).
67. Caporaso, J. G. *et al*. QIIME allows analysis of high-throughput community sequencing data. *Nature methods* **7**, 335–336, https://doi.org/10.1038/nmeth.f.303 (2010).
68. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* **6**, 65–70 (1979).
69. Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J. & Knight, R. UniFrac: an effective distance metric for microbial community comparison. *The ISME journal* **5**, 169–172, https://doi.org/10.1038/ismej.2010.133 (2011).
70. GOWER, J. C. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**, 325–338, https://doi.org/10.1093/biomet/53.3-4.325 (1966).
71. Anderson, M. J. & Willis, T. J. Canonical Analysis of Principal Coordinates: A Useful Method of Constrained Ordination for Ecology. *Ecology* **84**, 511–525, https://doi.org/10.1890/0012-9658(2003)084[0511:CAOPCA]2.0.CO;2 (2003).
72. Anderson, M. J. A new method for non-parametric multivariate analysis of variance. *Austral Ecol* **26**, 32–46, https://doi.org/10.1046/j.1442-9993.2001.01070.x (2001).
73. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550, https://doi.org/10.1186/s13059-014-0550-8 (2014).
74. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc B*, 289–300 (1995).
75. Grandhi, A., Guo, W. & Peddada, S. D. A multiple testing procedure for multi-dimensional pairwise comparisons with application to gene expression studies. *BMC bioinformatics* **17**, 104, https://doi.org/10.1186/s12859-016-0937-5 (2016).
76. Langille, M. G. *et al*. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature biotechnology* **31**, 814–821, https://doi.org/10.1038/nbt.2676 (2013).
77. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research* **40**, D109–114, https://doi.org/10.1093/nar/gkr988 (2012).
78. Calle, M. L., Urrea, V., Boulesteix, A. L. & Malats, N. AUC-RF: a new strategy for genomic profiling with random forest. *Human heredity* **72**, 121–132, https://doi.org/10.1159/000330778 (2011).

Acknowledgements

Research reported in this publication was supported in part by the US National Cancer Institute under award numbers R01CA159036, U01CA182370, R01CA164964, R03CA159414, P30CA016087, and R21CA183887. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. Samples were sequenced at the NYUMC Genome Technology Center. The NYUMC Genome Technology Center is partially supported by the Cancer Center Support Grant, P30CA016087, at the Laura and Isaac Perlmutter Cancer Center.

Author Contributions

B.A.P., J.A.S. and J.A. designed research; J.A.S., T.R.C., E.Y., C.F. and J.A. conducted research; B.A.P., R.B.H. and J.A. analyzed data; B.A.P., R.B.H. and J.A. wrote paper; G.M. and C.T. revised critically for content; B.A.P. and J.A. had primary responsibility for final content. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-28126-1.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.